>>> revista tradumàtica
Traducció i Tecnologies de la Informació i la Comunicació
Número 07: L'aplicació dels corpus lingüístics a la traducció : desembre 2009 :
http://www.fti.uab.cat/tradumatica/revista : ISSN: 1578-7559

La elaboración de un corpus *ad hoc* paralelo multilingüe[1]

Cristina Castillo Rodríguez
Departamento de Traducción e Interpretación
Universidad de Málaga
ccastillor@uma.es

Resumen
Este artículo presenta una propuesta de elaboración de un corpus *ad hoc* paralelo multilingüe, dividida claramente en una serie de fases. Por otro lado, se muestran las ventajas e inconvenientes a la hora de compilar un corpus de estas características, incidiendo sobre todo en la fase de alineación de los bitextos, imprescindible para este tipo de corpus.

Palabras clave
Lingüística de corpus, corpus *ad hoc*, corpus paralelo, traducción, alineación, bitexto

1. Introducción

Especialmente desde mediados del siglo XX numerosos estudios ya se estaban llevando a cabo bajo una metodología basada en corpus, a saber, la adquisición del lenguaje, la enseñanza de las lenguas extranjeras, la lingüística comparativa, la sintaxis y la semántica, entre otros. Sin embargo, la Lingüística de Corpus ha experimentado un renacimiento notable en los últimos años, en tanto ha ido ampliando su campo de aplicación, sobre todo, en lo que respecta al procesamiento del lenguaje natural, la ingeniería lingüística, la terminología y la traducción (cf. Corpas Pastor, 2003). Además, según Corpas Pastor, se trata más bien de una metodología de análisis aplicable a cualquier ámbito de la lingüística, ya que «se conforma como un nuevo paradigma de investigación más flexible y adaptable, que permite afrontar los retos actuales desde una multiplicidad de perspectivas» (Corpas Pastor, 2008: 49). Por eso, la Lingüística de Corpus es especialmente adecuada para la investigación de un fenómeno tan multidisciplinar como la Traducción.

Por otro lado, en las últimas décadas hemos asistido a una auténtica revolución tecnológica, la cual ha permitido la creación de corpus de mayor tamaño, gracias, sobre todo, a la red Internet, así como la gestión de grandes cantidades de textos, debido a la multitud de herramientas capaces de albergar y procesar corpus de textos en formato electrónico. Además, recientemente se ha reconocido el valor de los corpus en traducción como recursos útiles para el estudio lingüístico-contrastivo de las lenguas, lo cual ha contribuido sobremanera al rápido desarrollo de los estudios con corpus y, muy especialmente, de aquellos relacionados con la traductología, hasta llegar a crearse una corriente denominada Estudios de Traducción basados en Corpus (ETC); una corriente de la que, sin duda, su pionera es la autora Mona Baker (cf. Baker, 1993).

Así, para el estudio lingüístico-contrastivo de textos originales y sus traducciones, el corpus *ad hoc* se perfila como un recurso útil y eficaz que, además, dada la vasta variedad de recursos electrónicos y en red con los que contamos hoy en día, se puede elaborar de forma muy sencilla si se sigue una serie de pautas bien definidas, como las que se muestran a continuación.

2. El corpus *ad hoc* paralelo multilingüe

[1] El presente trabajo ha sido realizado en el seno del proyecto HUM-892 (Proyecto de Excelencia, Junta de Andalucía) y del proyecto *Ecosistema* (FF/2008-06080-C03-03/FILO, Ministerio de Ciencia y Tecnología)

En general, un corpus lingüístico es un conjunto, normalmente amplio, de textos o fragmentos de texto reales de una lengua determinada y que pueden tener origen escrito u oral. No obstante, aunque se han sucedido multitud de definiciones del término corpus (Baker, 1995; Bowker, 2002; Kenny, 2001; o Sinclair, 1991), la definición que podría considerarse más aceptada comúnmente es la que se propone en un documento técnico elaborado por EAGLES, acrónimo de *Expert Advisory Group on Language Engineering Standards* (1996), donde el corpus se define como «a collection of pieces of language that are selected and ordered according to explicit linguistic criteria in order to be used as a sample of the language».

El corpus *ad hoc* se trata de un tipo de recurso complementario que cubre ciertas necesidades de los usuarios que lo compilan. Los corpus *ad hoc* (denominación acuñada por Aston, 1999) han recibido distintas denominaciones, a saber, *virtual corpus* (Ahmad et al., 1994); *special purpose corpus* (Pearson, 1998); *corpus especials* (Sánchez-Gijón, 2003); *customized corpus* (Austermülh, 2001); *disposable* (Varantola, 2000); *do-it-yourself (DIY) corpora* (Zanettin, 2002), entre otras. El corpus *ad hoc* se basa en la compilación de textos a través de la red. Además, según Corpas Pastor (2002: 195), cuando se compila un corpus de este tipo no se trata sólo de solventar un problema puntual sino de «reunir toda la documentación disponible sobre un tema en muy poco tiempo».

Por otro lado, a partir de los años noventa, los corpus paralelos constituyen un campo de trabajo de especial interés tanto en los ETC como dentro de la Lingüística Computacional. En general, los textos contenidos en un corpus paralelo «are natural-language texts that have the same semantic content, but are expressed in different forms» (Nevill y Bell, 1992: 3). De hecho, un corpus paralelo está compuesto por textos originales y sus traducciones a otra lengua. Es por ello que suelen recibir la denominación de corpus paralelo bilingüe o bitexto, término acuñado por Harris (1988: 8) quien, desde su punto de vista, considera que un texto original y su traducción no son en realidad dos textos sino que conforman «a single text in two dimensions, each of which is a language».

Además, la compilación de un corpus *ad hoc* paralelo bilingüe o multilingüe puede ser de especial utilidad por dos razones fundamentales: de un lado, como diccionario basado en contextos reales de un dominio de especialidad determinado, y, de otro, como fuente de datos para el análisis contrastivo y posterior evaluación de traducciones en un campo de especialidad.

Como ejemplo de diccionario basado en contextos reales destacamos la compilación de un corpus *ad hoc* paralelo multilingüe en el campo de la farmacología, en concreto, en las fichas técnicas de productos sanitarios para diagnóstico *in vitro*, cuyos textos, además de haber sido traducidos por traductores, han sido revisados por un comité de expertos en la materia (cf. Castillo Rodríguez, 2005 y 2007).

Como fuente de información para llevar a cabo un estudio contrastivo entre las distintas versiones traducidas a varias lenguas de un mismo texto original, subrayamos el ejemplo que nos proponemos describir en cada una de las fases de elaboración de un corpus *ad hoc* paralelo multilingüe, es decir, textos procedentes del ámbito del turismo, en especial, los que están relacionados con el segmento del turismo denominado salud y belleza, escritos originalmente en español y traducidos a las lenguas inglesa, italiana y francesa.

3. Compilación de un corpus *ad hoc* paralelo multilingüe

La fase de creación del corpus es esencial en cualquier estudio que pretenda basarse en este recurso. Por tanto, será fundamental seguir una metodología de compilación del corpus, así

como las decisiones que se adopten en cuanto a los criterios de diseño, previos a la compilación en sí.

A continuación, esbozaremos una serie de fases para poder crear un corpus *ad hoc* paralelo multilingüe de textos de un dominio concreto. Estas fases las hemos dividido en cuatro, es decir, la fase de recopilación de los datos, la fase de almacenamiento, la fase de conversión de formatos y la fase de alineación.

3.1. Fase de recopilación de datos

En la actualidad, el uso de los corpus *ad hoc* supone un gran avance para el manejo y extracción de información a partir de la documentación que puede descargarse de la red Internet. La selección de los documentos que integrarán un corpus *ad hoc* paralelo, sin embargo, no puede ser aleatoria, sino que debe atender a unos criterios concretos de diseño para conseguir un grado de representatividad lo más adecuado posible. Es por ello que dentro de esta fase de recopilación de los datos podemos distinguir dos subtareas: por un lado, la elaboración de los criterios de diseño y, por otro, la búsqueda de la información.

3.1.1. Criterios de diseño

Aunque no hay consenso a la hora de describir los criterios que han de ser tenidos en cuenta para el diseño de los corpus, sí lo hay en que se deben especificar unos criterios de diseño previamente a la compilación en sí de los datos para que el corpus, en este caso, ad hoc, sea representativo de un campo del conocimiento en concreto. Además, los criterios que se emplean para diseñar un corpus dependerán de las necesidades investigadoras para las cuales se compila un corpus determinado. Por ello, y basándonos en los criterios de diseño elaborados por Bowker y Pearson (2002: 54), proponemos los siguientes: i) el propósito de la compilación; ii) el tamaño; iii) el medio; iv) la temática; v) el tipo textual; vi) la autoría; vii) la fecha de publicación; y, por último, viii) la/s lengua/s.

A modo de ejemplo, consideraremos que el *propósito de la compilación* de un corpus *ad hoc* paralelo multilingüe es la investigación sobre la calidad de la traducción de los textos publicados en la red. Mucho se ha discutido acerca de cuál debe ser el *tamaño* de un corpus compilado para un fin específico. Autores como Bowker y Pearson (2002) afirman que no hay reglas fijas que puedan establecer el tamaño ideal de un corpus, mientras que otros, como Kennedy (1998), están de acuerdo en que un corpus muy extenso no tiene que ser necesariamente más útil que uno de menor tamaño, ya que la utilidad dependerá de la finalidad de la compilación y, sobre todo, de la representatividad de los textos en un ámbito determinado. Además, como advierte Corpas Pastor (2001: 165): «conviene recordar que la compilación de un corpus *ad hoc* viene determinada por la necesidad perentoria de acceder a documentación específica sobre un tema concreto a la mayor brevedad posible». Por otro lado, diversos estudios realizados recientemente han demostrado que un corpus de textos de un determinado tipo textual no tiene porqué ser muy extenso para que sea representativo. De hecho, Corpas Pastor y Seghiri Domínguez (2006; 2007a; 2007b; 2008/en prensa) han desarrollado un método para calcular el umbral mínimo de representatividad de un corpus mediante el algoritmo N-Cor de análisis de la densidad léxica en función del aumento incremental del corpus. Se trata de una solución eficaz para determinar *a posteriori*, por primera vez de forma objetiva y cuantificable, el tamaño mínimo que debe alcanzar un corpus para que pueda ser considerado representativo en términos estadísticos; un método que se ha visto implementado en la aplicación ReCor, que en breve estará disponible online.

El *medio* será escrito, aunque a este respecto añadiremos que se tratan de textos escritos digitales exclusivamente en línea, ya que en ningún momento se incluirán documentos en

formato papel o textos escaneados. En cuanto a la *temática*, nos centraremos principalmente en el segmento del turismo denominado turismo de salud y belleza, aunque dentro de la *tipología textual* encontrada en tal segmento descargaremos únicamente aquellos textos relacionados con el material promocional del mismo, el cual está compuesto por todos aquellos folletos turísticos del dominio en cuestión.

La *autoría* se considera otro de los criterios a tener en cuenta, ya que siempre se deben compilar textos cuyos autores muestren cierta fiabilidad. En nuestro caso, al tratarse de un segmento en particular del turismo, nos centraremos en páginas web de organismos institucionales, así como de cadenas hoteleras conocidas. Por lo que respecta a la *fecha de publicación*, se recomienda la inclusión de textos actuales, por lo que, para el ejemplo que mostramos, habría que tener en cuenta la fecha de actualización de las páginas que, en general, viene reflejada en la página de inicio de las mismas. Además, otro de los indicios de actualidad de una página la encontramos en las promociones que se ofertan, cuyas fechas de inicio y finalización nos muestran la actualización de las páginas en sí. Por último, las *lenguas* seleccionadas para este estudio con corpus *ad hoc* paralelo multilingüe son el español, como lengua original, así como el francés, inglés e italiano, como lenguas meta.

3.1.2. Búsqueda de la información

No hay duda de que hoy en día existe una gran diversidad de fuentes de información que se encuentran disponibles en la red; no obstante, autores como Zanettin advierten de que existen diversos problemas en relación con el uso de los documentos extraídos de este medio. En concreto el autor afirma que hay dos problemas principales (2002: 241):

> The first concerns procedures for assessing relevance and reliability: Information is dispersed in the WWW through vast quantities of documents, and it is thus crucial for the translator to retrieve this information in the most efficient and effective way. The second relates to strategies and techniques for searching electronic texts: Search engines provide access points to Internet documents either through lists generated by full text searches or by pre-selected lists organized by topic, and are thus catalogues rather than corpora.

En este sentido, Austermühl (2001: 52 y ss.), aunque reconoce que encontrar información fiable puede resultar una tarea bastante difícil, afirma que la búsqueda de datos generales online no tiene por qué plantear demasiados problemas. Para evitar perderse en el ciberespacio, se necesitan unas estrategias de búsqueda bien definidas. Por ello, nos centraremos en dos de tipos de búsqueda propuestos por Austermühl, esto es, en las búsquedas institucionales y en las búsquedas por palabras clave.

Entre las búsquedas institucionales, para extraer los datos relacionados con el tema que nos ocupa, destacamos, por ejemplo, la Organización Mundial del Turismo (OMT) o la Consejería de Turismo, Comercio y Deporte, de la Junta de Andalucía. Por otro lado, estamos de acuerdo con Merlo Vega (2004: 315) en que las fuentes de información institucionales: «son aquellas que aportan información sobre empresas y entidades, ya sean históricas, de localización, de estructura, de su actividad o de cualquier otro tipo». Por tanto, consideramos conveniente la búsqueda a través de cadenas hoteleras, de turoperadores, así como de agencias de viajes que estuvieran en línea (como, por ejemplo, Viajes Marsans, Halcón Viajes, Viajes El Corte Inglés, así como Accorhotels.com, Barceló Hotels & Resorts, entre otros).

La búsqueda por palabras clave se trata del procedimiento más empleado puesto que suele ser más flexible, aunque cuenta con el inconveniente de mostrar un exceso de documentos no pertinentes para el corpus que se desea crear, es decir, presenta demasiado «ruido» documental, por lo que habrá opciones que no nos interesen. Para acotar los resultados, se suelen utilizar técnicas de búsqueda más complejas, como es el caso de los operadores booleanos o truncamientos.

Aunque existe una vasta variedad de buscadores y metabuscadores, nuestras búsquedas se llevarán a cabo a través de uno de los buscadores online mundialmente conocido; es el caso del motor de búsqueda Google, el cual, de acuerdo con un gran número de analistas, supone, en la actualidad, el mejor motor de búsqueda en cuanto a la calidad que ofrecen los resultados obtenidos (cf. Radev et al., 2005). Mediante las opciones avanzadas de este buscador, debemos seleccionar: i) la lengua, esto es, español; ii) la opción «páginas de España», con el fin de evitar la presencia de documentos provenientes de otros países en los que se hable también español (por ejemplo, Chile, México, Argentina, etc.); iii) las distintas técnicas de búsqueda booleanas para la extracción de los textos especializados, como por ejemplo, AND, OR, NOT (por ejemplo, una ecuación de búsqueda sería spa OR balneario AND turismo AND hotel).

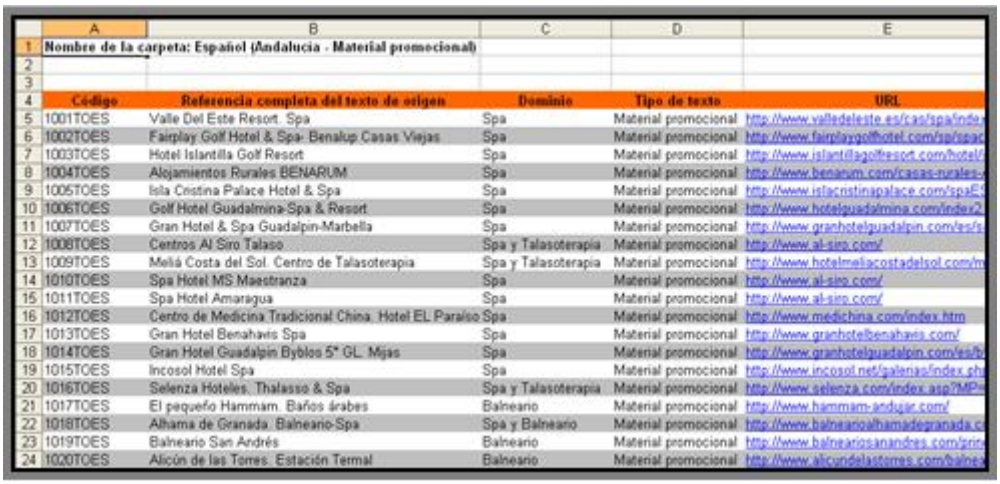
3.2. Fase de almacenamiento

El siguiente paso dentro del protocolo de compilación de un corpus es, precisamente, el almacenamiento de la información, que contempla, además de un almacenamiento básico, el tipo de codificación que se ha utilizado para cada uno de los registros compilados en la red.

Para almacenar los datos se necesita una codificación que sea unívoca de cada uno de los registros compilados para poder proceder a la gestión y análisis posteriores. Por tanto, de forma simultánea al almacenamiento de los textos seleccionados, debemos asignar una codificación que sea específica, de forma que puedan localizarse y extraerse posteriormente cada uno de los registros de los tres bitextos compilados en el corpus paralelo, a saber, el bitexto español-inglés, español-italiano y español-francés.

Por otro lado, al mismo tiempo que se asigna la codificación específica para cada uno de los registros, recomendamos la elaboración de una tabla, en un archivo creado a través del programa *Excel* de *Microsoft Office*, en la que se muestre un breve resumen de los registros que formarán el corpus. La información contenida en dicha tabla, dividida en cinco columnas, contemplará los siguientes apartados: el «Código» del registro; otra columna con la «Referencia completa del texto de origen», en la que introducimos el título de la página web en cuestión; otra que describe el «Dominio», donde especificamos si un texto contiene información procedente de un spa, balneario o centro de talasoterapia, o bien un híbrido entre las tres; una columna para el «Tipo de texto», en la que sólo se menciona que es material promocional; y una última columna con la «URL» desde donde se ha descargado la información.

Dentro de los registros originales en lengua española se ha establecido una serie de códigos cifrados entre números y letras, cuyos registros se comprenden entre 1001 hasta 1999, en el que el 1 millar es el número que corresponde a España, es decir, que los textos compilados y volcados a esta carpeta sólo se habrán compilado en España. A estos números les siguen las siglas TOES, donde «TO» atiende a texto original y «ES» a lengua española. A modo de ilustración, mostramos la siguiente tabla en *Excel* con la referencia de los veinte primeros registros compilados de páginas web en materia del segmento del turismo de salud y belleza:



	A	B	C	D	E
1	Nombre de la carpeta: Español (Andalucía - Material promocional)				
2					
3					
4	Código	Referencia completa del texto de origen	Dominio	Tipo de texto	URL
5	1001TOES	Valle Del Este Resort. Spa	Spa	Material promocional	http://www.valledeleste.es/cas/spa/inde
6	1002TOES	Fairplay Golf Hotel & Spa- Benalup Casas Viejas	Spa	Material promocional	http://www.fairplaygolfhotel.com/sp/spac
7	1003TOES	Hotel Islantilla Golf Resort	Spa	Material promocional	http://www.islantillagolfresort.com/hotel/
8	1004TOES	Alojamientos Rurales BENARUM	Spa	Material promocional	http://www.benarum.com/casas-rurales-
9	1005TOES	Isla Cristina Palace Hotel & Spa	Spa	Material promocional	http://www.islacristinapalace.com/spa/E
10	1006TOES	Golf Hotel Guadalmina Spa & Resort	Spa	Material promocional	http://www.hotelguadalmina.com/index2
11	1007TOES	Gran Hotel & Spa Guadalpin-Marbella	Spa	Material promocional	http://www.granhotelguadalpin.com/es/s
12	1008TOES	Centros Al Siro Talaso	Spa y Talasoterapia	Material promocional	http://www.al-siro.com/
13	1009TOES	Meliá Costa del Sol. Centro de Talasoterapia	Spa y Talasoterapia	Material promocional	http://www.hotelmeliacostadelsol.com/m
14	1010TOES	Spa Hotel MS Maestranza	Spa	Material promocional	http://www.al-siro.com/
15	1011TOES	Spa Hotel Amaragua	Spa	Material promocional	http://www.al-siro.com/
16	1012TOES	Centro de Medicina Tradicional China. Hotel EL Paraíso	Spa	Material promocional	http://www.medichina.com/index.htm
17	1013TOES	Gran Hotel Benahavís Spa	Spa	Material promocional	http://www.granhotelbenahavis.com/
18	1014TOES	Gran Hotel Guadalpin Byblos 5* GL. Mijas	Spa	Material promocional	http://www.granhotelguadalpin.com/es/b
19	1015TOES	Incosol Hotel Spa	Spa	Material promocional	http://www.incosol.net/galenas/index.ph
20	1016TOES	Selenza Hoteles. Thalasso & Spa	Spa y Talasoterapia	Material promocional	http://www.selenza.com/index.asp?MP=
21	1017TOES	El pequeño Hammam. Baños árabes	Balneario	Material promocional	http://www.hammam-andujar.com/
22	1018TOES	Alhama de Granada. Balneario-Spa	Spa y Balneario	Material promocional	http://www.balnearioalhamadegranada.c
23	1019TOES	Balneario San Andrés	Balneario	Material promocional	http://www.balneariosanandres.com/prin
24	1020TOES	Alicún de las Torres. Estación Termal	Balneario	Material promocional	http://www.alicundelastorres.com/balnea

Tabla 1. *Muestra de los registros en español compilados de páginas web*

Para los registros de textos traducidos en lengua inglesa, italiana y francesa se ha procedido a la asignación de una nueva codificación: 1001TMEN hasta 1999TMEN, 1001TMIT hasta 1999TMIT y 1001TMFR hasta 1999TMFR, donde «TM» atiende a texto meta, mientras que «EN», «IT», «FR», se corresponden con cada una de las lenguas, esto es, inglesa, italiana y francesa. De esta forma, 1001TMEN, 1001TMIT y 1001TMFR se corresponden con su registro original 1001TOES y así sucesivamente con todos los registros en los que se haya encontrado traducción publicada en las lenguas citadas.

Todos los registros con sus códigos unívocos se guardarán por lenguas en carpetas independientes.

3.3. Fase de conversión de formatos

Para poder analizar los textos del corpus específico que se han seleccionado y guardado es necesario transformarlos en un formato que puedan reconocer los programas de gestión de corpus. Los textos descargados de la red suelen estar codificados en una gran variedad de formatos, aunque en la mayor parte de ellos aparecen en formato HTML o en formato PDF. Por tanto, los textos deben convertirse a un formato de texto plano (extensión .txt), que es el formato requerido por la mayoría de programas de concordancias y de gestión de corpus. Mientras que para la conversión de los textos en formato HTML a texto plano sólo se requiere volver a guardar los archivos con la nueva extensión, para convertir los textos codificados en formato PDF se precisa el uso de programas concretos —de acceso libre,aunque también nos hemos valido de versiones *demo* de algunos programas—, que permiten la conversión de estos tipos de textos a formato plano (como, por ejemplo, el programa pdf2txt).

Una vez que se han transformado a este nuevo formato en el que el texto aparece sin cursivas, negritas, etc., hace falta «limpiarlos», ya que en muchas ocasiones el proceso de transformación a texto sin formato, o formato de texto plano, provoca la aparición de errores de conversión. En cuanto a los errores que podemos encontrar a la hora de convertir un texto con extensión .pdf a texto plano, señalamos, por ejemplo, la adición de tildes (o la ausencia de éstas), la adición de otros códigos no propios del texto original y el no reconocimiento de ciertas letras como es el caso de la «ñ», entre otros aspectos. En estos casos, habría que comprobar la fuente del texto para saber a qué se refieren los nuevos códigos y proceder a reemplazarlos por la letra o palabra acentuada correcta. Por otro lado, los archivos HTML suelen presentar palabras y terminologías pertenecientes al vocabulario propio del soporte web, tales como «Aceptar», «Volver», «Home», entre otras, las cuales habría que eliminarlas

para que no computen en el recuento total de palabras totales del corpus, ya que no son representativas para el tipo de estudio que nos proponemos llevar a cabo.

Para la limpieza de los textos es necesario el uso de un procesador de textos, por ejemplo, el programa *Word* de *Microsoft Office*, cuya opción «buscar y reemplazar» pueda facilitar la tarea de eliminación y corrección de las palabras «mal convertidas». Así, procederemos a un nuevo almacenamiento de los textos con los mismos códigos asignados anteriormente, aunque esta vez en formato RTF, el cual nos permitirá tratar los bitextos en la siguiente fase de alineación, una tarea que se llevará a cabo mediante la herramienta *WinAlign* del programa *TRADOS*.

3.4. Fase de alineación de los bitextos

A efectos de investigación, no basta con la recopilación de textos para formar un corpus paralelo, puesto que para poder aprovechar al máximo todos los datos compilados, éstos deben estar estructurados de manera tal que se facilite su explotación y posterior análisis lingüístico-contrastivo. Uno de los tratamientos más importantes que se aplica al corpus paralelo es lo que se conoce como alineación, puesto que es «el proceso que mayor valor añadido aporta a un corpus multilingüe» (Abaitua, 2002: 6). De este modo, este proceso constituye una de las fases más importantes y necesarias dentro del protocolo de compilación de un corpus paralelo, que generará lo que se conoce comúnmente como corpus paralelo alineado.

El proceso de alineación de un corpus paralelo consiste en la reestructuración de los textos de tal forma que se pueda establecer una correspondencia entre los párrafos, oraciones y/o palabras de los textos involucrados. Para llevar a cabo el proceso de alineación de los textos originales y los textos traducidos, contamos en la actualidad con varias herramientas que permiten realizar este proceso. De hecho, el programa de gestión de corpus *ParaConc* contiene, además de módulos de análisis para el estudio contrastivo de textos de hasta cuatro lenguas simultáneamente, un módulo de alineación sencilla de textos.

Por otro lado, para alineaciones más complejas, esta herramienta requiere que los textos se procesen con otro alineador automático como, por ejemplo, *WinAlign*, un módulo de análisis contenido dentro del paquete de herramientas *TRADOS*, que destaca fundamentalmente por presentar un editor interactivo y flexible de la alineación. Este módulo propone una serie de alineaciones que el usuario debe ir validando de forma manual. Una vez que se modifican y se validan las secuencias alineadas, permite la exportación de los resultados en un archivo de texto plano, formato requerido por la mayoría de programas de gestión de corpus actuales. A continuación, ejemplificamos uno de los textos alineados con *WinAlign*:

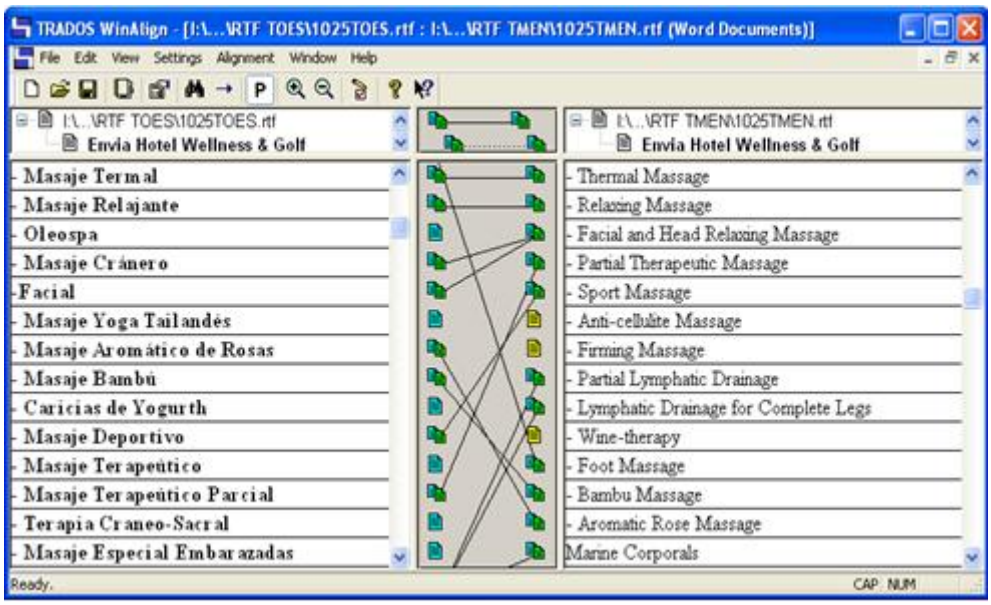


Ilustración 1. *Ejemplo de alineación de un bitexto con WinAlign*

4. Ventajas e inconvenientes

Hoy en día, los corpus *ad hoc* paralelos multilingües constituyen una gran fuente de información documental y terminológica al servicio de los traductores e investigadores interesados en estudios y análisis contrastivos entre lenguas. La principal ventaja que podemos subrayar a la hora de crear este tipo de corpus reside sobre todo en la disponibilidad en la red de multitud de textos traducidos a varias lenguas, así como en el tiempo en que se puede recopilar tal información.

Sin embargo, un corpus de textos no constituye en sí mismo un recurso útil, puesto que, en palabras de Leech (1991: 22): «it is widely acknowledged today that a corpus needs the support of a sophisticated computational environment, providing software tools both to retrieve data from the corpus and to process linguistically the corpus itself». De hecho, hoy en día, contamos con una gran variedad de recursos y herramientas informáticas capaces de procesar y analizar aspectos lingüísticos, fraseológicos y discursivos de textos escritos en una lengua determinada, así como de esos mismos textos traducidos a más de una lengua. Entre estas herramientas destacan los denominados programas de gestión de corpus, que constituyen una gran ayuda para el traductor o lingüista, ya que permiten analizar de forma rápida la frecuencia de aparición de una determinada palabra en un contexto dado, tanto en el texto original (TO) como en el texto meta (TM). Estos programas pueden descargarse directamente de la red, diferenciándose unos de otros en las utilidades que ofrecen (cf. Castillo Rodríguez, 2005, 2007, y Corpas Pastor, 2007, donde se describen algunos de estos programas).

El principal inconveniente estriba en la alineación de los bitextos cuando se trata de un corpus paralelo. Rabadán y Fernández Nistal (2002: 76-77) afirman que esta tarea puede realizarse de forma manual en el caso de textos de pequeña extensión, aunque, por otro lado, reconocen que puede resultar demasiado laboriosa cuando se trata de alinear grandes corpus de textos, para lo cual se debe recurrir a herramientas informáticas que permitan la alineación automática de los textos.

Por otro lado, la dificultad de la alineación automática de textos reside principalmente en la organización de los segmentos paralelos tanto en el TO como en el TM. El grado de dificultad

de alineación de textos variará dependiendo del tipo de texto. Por ejemplo, en textos como las fichas técnicas de ciertos productos de farmacología, la organización textual en los TM suele ser similar a la del TO, ya que estos tipos de textos suelen mostrar una serie de convenciones textuales más estrictas. Como consecuencia, la alineación de tales textos puede resultar una tarea bastante sencilla; un proceso podría llevarse a cabo mediante el módulo de alineación contenido en *ParaConc*.

Por el contrario, cuando se trabaja con textos de carácter publicitario como el que presentamos en esta metodología de elaboración de corpus *ad hoc* paralelo multilingüe, la tarea de alineación resulta más complicada. Los textos relacionados con el material promocional de un determinado segmento turístico, por lo general, suelen estar ligados a cuestiones de marketing, los cuales dan lugar a constantes variaciones en el material publicado y, por tanto, en el orden de aparición de determinadas secuencias en cada una de las lenguas en las que se presenta dicho material, además de eliminarse determinados servicios en algunos de los textos traducidos. Es por ello que se deben emplear herramientas de alineación de carácter flexible que permitan la alineación y/o validación de secuencias originales y traducidas, como hemos descrito anteriormente.

5. Conclusiones

Es un hecho que hoy en día los corpus se están empleando en multitud de campos de aplicación y, entre ellos, los Estudios de Traducción basados en Corpus. El corpus *ad hoc* paralelo multilingüe constituye un recurso eficaz, no sólo como fuente documental de traducciones basadas en contextos reales, sino para llevar a cabo estudios contrastivos entre lenguas.

Pero, previamente al análisis de grandes cantidades de textos, es necesario seguir un protocolo de elaboración del corpus en sí. Para ello, hemos descrito de forma detallada una serie de fases que nos llevarán a la creación de un corpus lo más representativo posible de un dominio concreto. Sin embargo, a pesar de las ventajas que nos brindan las nuevas tecnologías que están a disposición de cualquier usuario, la elaboración de este tipo de corpus conlleva una serie de inconvenientes, como es el caso de la fase de alineación de segmentos, a la que se debe prestar especial importancia, puesto que sin ella, el análisis contrastivo no podría llevarse a cabo con éxito.

Bibliografía

Abaitua Odriozola, J. (2002). «Tratamiento de corpora bilingües». En Martí Antonín, M.A. y J. Llisterri Boix (eds). *Tratamiento del lenguaje natural*. Barcelona: Universidad Autónoma de Barcelona. 61-90.

Ahmad, K.; P. Holmes-Higgin y S. Raza Abidi. (1994). «A description of texts in a corpus: 'Virtual' and 'real' corpora». En Martin, W.; W. Mejis; M. Moerland; E. ten Pas; P. van Sterkenburg y P. Vossen (eds.). *EURALEX 1994: Proceedings. Papers submitted to the 6th Euralex International Congress on Lexicography*. Ámsterdam: Vrije Universiteit. 390-402.

Aston, G. (1999). «Corpus use and learning to translate». *Textus,* 12.

Austermühl, F. (2001). *Electronic Tools for Translators*. Manchester: St. Jerome.

Baker, M. (1993). «Corpus Linguistics and Translation Studies: implications and applications». En Baker, M., Francis, G. & E. Tognini-Bonelli (eds.) *Text and Technology: In Honour of John Sinclair*. Ámsterdam y Filadelfia: John Benjamins. 233-250.

Baker, M. (1995). «Corpora in Translation Studies: An Overview and Some Suggestions for Future Research». *Target. International Journal of Translation Studies*. 7 (2). Ámsterdam: John Benjamins Publishing Company. 223-243.

Bowker, L. (2002). *Computer-Aided Translation Technology. A practical introduction*. Ottawa: University of Ottawa Press.

Bowker, L. y J. Pearson. (2002). *Working with Specialized Language: A practical guide to using corpora.* Londres: Routledge.

Castillo Rodríguez, C. (2005). *La compilación de un corpus paralelo bilingüe (inglés-español) para el análisis de los aspectos terminológicos de los productos sanitarios para diagnóstico* in vitro. Trabajo de investigación. Málaga: Universidad de Málaga. [Sin publicar]

Castillo Rodríguez, C. (2007). «A compilation of a bilingual parallel corpus (English-Spanish) specialized in *in vitro* diagnostic medical devices (IVD)». En *Actas del X Simposio Internacional de Comunicación Social: 20 años de comunicación científica*. Santiago de Cuba: Centro de Lingüística Aplicada. Ministerio de Ciencia, Tecnología y Medio Ambiente. 677-681.

Corpas Pastor, Gloria. (2001). «Compilación de un corpus *ad hoc* para la enseñanza de la traducción inversa especializada». *TRANS. Revista de Traductología*, 5. 155-184.

Corpas Pastor, G. (2002). «Traducir con corpus: de la teoría a la práctica». En García Palacios, J. y M.T. Fuentes Morán (eds.). *Texto, Terminología y Traducción*. Salamanca: Almar. 189-226.

Corpas Pastor, G. (2003). «Turicor: Compilación de un corpus de contratos turísticos (alemán, español, inglés, italiano) para la generación textual multilingüe y la traducción jurídica». En Ortega Arjonilla, E. (dir.); E. Echeverría Pereda, E. Alarcón Navío y C. Mata Pastor (coords.). *Panorama actual de la investigación en traducción e interpretación*. Vol. II. Granada: Atrio. 373-384.

Corpas Pastor, G. (2007). «Lost in Specialised Translation: The Corpus as an Inexpensive and Under-Exploited Aid for Language Service Providers». *Translating and the Computer 29. Proceedings of the ASLIB Conference*. Londres: Aslib. 1-18.

Corpas Pastor, G. (2008). *Investigar con corpus en traducción: los retos de un nuevo paradigma*. Frankfurt: Peter Lang.

Corpas Pastor, G. y M. Seghiri Domínguez. (2006). *El concepto de representatividad en lingüística de corpus: aproximaciones teóricas y consecuencias para la traducción*. Documento Técnico. Departamento de Traducción e Interpretación. Universidad de Málaga. [BFF2003-04616 MCYT/TI-DT-2006-1].

Corpas Pastor, G. y M. Seghiri Domínguez. (2007*a*). «Determinación del umbral de representatividad de un corpus mediante el algoritmo N-Cor». *Procesamiento del Lenguaje Natural*, 39. 165-172.

Corpas Pastor, G. y M. Seghiri Domínguez. (2007*b*). «Specialized Corpora for Translators: A Quantitative Method to Determine Representativeness». *Translation Journal*. 11. (3).

Corpas Pastor, G. y M. Seghiri Domínguez. (2008/en prensa). *El concepto de representatividad en lingüística de corpus: aproximaciones teóricas y consecuencias para la traducción*. Málaga: Servicio de Publicaciones de la Universidad.

Eagles (Expert Advisory Group on Language Engineering Standards). (1996). «Text corpora Working Group reading Guide». *EAGLES Document EAG-TCWG-FR-2.* Versión de mayo de 1996. S. pag. < http://www.ilc.cnr.it/EAGLES/corpintr/corpintr.html >

Harris, B. (1988). «Bi-text, a New Concept in Translation Theory». *Working Papers on Bilingualism* o *Language Monthly*, 54. 8-10.

Kennedy, G. (1998). *An Introduction to Corpus Linguistics*. Londres; Nueva York: Longman.

Kenny, D. (2001). *Lexis and Creativity in Translation. A Corpus-based Study*. Manchester y Northampton: St. Jerome.

Leech, G. (1991). «The state of the art in corpus linguistics». En Aijmer, K. y B. Altenberg. (eds.). *English Corpus Linguistics*. Londres: Longman. 8-29.

Merlo Vega, J.A. (2004). «Uso de la documentación en el proceso de traducción especializada». En Gonzalo García, C. y V. García Yebra (eds.). *Manual de documentación y terminología para la traducción especializada*. Madrid: Arco/Libros. 309-336.

Nevill, C. y T. Bell. (1992). «Compression of parallel texts». *Information Processing and Management*, 28 (6). Gran Bretaña: Pergamon Press. 781-793.

Pearson, J. (1998). *Terms in Context, Studies in Corpus Linguistics*, 1. Ámsterdam y Filadelfia: John Benjamins.

Rabadán Álvarez, R. y P. Fernández Nistal. (2002). *La traducción inglés-español: fundamentos, herramientas, aplicaciones.* León: Servicio de publicaciones de la Universidad de León.

Radev, D.; W. Fan; H. Qi; H. Wu y A. Grewal. (2005). «Probabilistic question answering on the web». *Journal of the American Society for Information Science and Technology (JASIST)*, 56 (6). 571-583.

Sánchez-Gijón, P. (2003). «És la web pública la nova biblioteca del traductor?». *Tradumàtica: Traducció i tecnologies de la informació i la comunicación*, 2.

Sinclair, J. (1991). *Corpus, Concordance, Collocation*. Oxford: Oxford University Press.

Varantola, K. (2000). «Translators, dictionaries and text corpora». En Bernardini, S. y F. Zanettin (eds.). *I corpora nella didattica della traduzione*. Bolonia: CLUEB. 117-133.

Zanettin, F. (2002). «DIY Corpora: The WWW and the Translator». En Belinda Maia; Jonathan Haller y Margherita Urlrych (eds.). *Training the Language Services Provider for the New Millennium.* Oporto: Facultade de Letras, Universidade do Porto. 239-248.